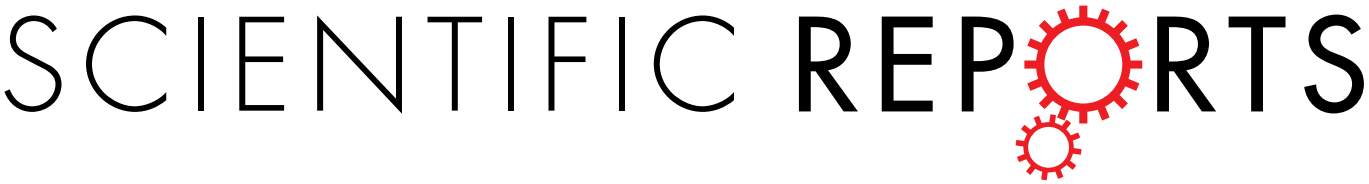
Inference of causality in epidemics on temporal contact networks



OPEN

Inference of causality in epidemics on temporal contact networks

Alfredo Braunstein[1,2,3,*] & Alessandro Ingrosso[1,*]

Received: 01 March 2016
Accepted: 11 May 2016
Published: 10 June 2016

Investigating into the past history of an epidemic outbreak is a paramount problem in epidemiology. Based on observations about the state of individuals, on the knowledge of the network of contacts and on a mathematical model for the epidemic process, the problem consists in describing some features of the posterior distribution of unobserved past events, such as the source, potential transmissions, and undetected positive cases. Several methods have been proposed for the study of these inference problems on discrete-time, synchronous epidemic models on networks, including naive Bayes, centrality measures, accelerated Monte-Carlo approaches and Belief Propagation. However, most traced real networks consist of short-time contacts on continuous time. A possibility that has been adopted is to discretize time line into identical intervals, a method that becomes more and more precise as the length of the intervals vanishes. Unfortunately, the computational time of the inference methods increase with the number of intervals, turning a sufficiently precise inference procedure often impractical. We show here an extension of the Belief Propagation method that is able to deal with a model of continuous-time events, without resorting to time discretization. We also investigate the effect of time discretization on the quality of the inference.

Identifying past features of an epidemic outbreak remains a challenging problem even for simple stochastic epidemic models, such as the susceptible-infected (SI) model and the susceptible-infected-recovered (SIR) model. In recent years, this problem has received considerable attention, especially on discrete time models[1–5]. For these models, we recently proposed an approximate Bayesian method based on Belief Propagation (BP)[6,7], that gave the first exact tractable solution to a family of discrete time inference problems on acyclic graphs and an excellent approximation on general graphs, including real ones. The problem addressed ranged from the inference of the epidemic source (the patient zero or index case), inference of the infection times and the epidemic parameters, all from the knowledge of the network plus a (partial, noisy) snapshot of the infection state of the system at a given instant.

In the last years, several precise spatio-temporal information about contacts between individuals in a community have been collected, representing close proximity[8,9], social or sexual interactions[10,11] and more. Each dataset consists of a time-stamped list of pairs of individuals. Seeking to explore characteristics of potential outbreaks, many authors studied the disease propagation over those communities employing compartment infection models such as SI and SIR. Technically, a simple way to achieve this is by computing a weighted discrete time network. This can be done by sub-dividing the time line into subintervals of length Δ (time-steps), aggregating all contacts falling in a given interval $[t\Delta,(t+1)\Delta]$ into a time-step dependent weight λ_{ij}^{t} equal to $1-(1-\lambda)^{k_t}$, where λ is the probability of transmission in a single quasi-instantaneous contact and k_t the number of contacts the interval[8,6]. Once this discrete time network has been constructed, the spread of infectious diseases on the community can be described through a discrete time SIR model, in which the transition probabilities between states defining each of these models depend on the time-step t. However, these coarsening methods naturally lead a loss of timing information and precision, becoming exact only in the limit of small Δ and a large number of intervals. Unfortunately, the computational time of both simulations and various inference algorithms typically increase with the number of time steps, making a sufficiently precise analysis unpractical, if not impossible in most cases. In the following, we will describe a very simple semi-continuous time stochastic model of infection dynamics that does not require coarsening or binning and is naturally equivalent to the $\Delta \to 0$ limit. For simplicity, we will concentrate on the SIR model, but all methods here can be naturally generalized to other variants such as SEIR[8]. We will then develop a semi-continuous time inference framework which is able to deal with contact network datasets without any discretization approximation, may it be implicit or explicit. The method

[1]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. [2]Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy. [3]Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.B. (email: alfredo.braunstein@polito.it)

will be shown to perform very well on two datasets of real contact networks, being able to reconstruct the epidemic source, the infection times and the infection causality tree with a great degree of accuracy in a wide range of parameters. In the concluding section, its performance will be compared to a discretized version of the model, showing that the inference performance under discretization approximations, although converging to the non-discretized one, does so in a range of discretization precision that renders it extremely unpractical.

Methods

A static model to describe dynamics. Let us consider an evolving contact network G composed of N nodes. Each node i is equipped with a time dependent state variable $x_i(t) \in \{S, I, R\}$ so that at time t it can be in one of three possible states: susceptible (S), infected (I), and recovered/removed (R). We will define our dynamical model as follows: let time $t \in \mathbb{R}_+$ be continuous, and contagion events be instantaneous with probability λ. Each pair of individuals (i, j) will be in contact in a discrete set of instants $T_{ij}(0) < T_{ij}(1) < \ldots < T_{ij}(n_{ij})$ (given by the real network traced dataset), where we assume for simplicity $T_{ij}(r) = \infty$ for $r > n_{ij}$. As time advances, contagion to node j will happen at a time t_j with probability λ if i is infected, j is susceptible and there exists a contact between the two nodes with some index r_i such that $T_{ij}(r_i) = t_j$. Let us define $t_i = \min\{t: x_i(t) = I\}$ the time at which node i gets infected (infection time) and $g_i = \min\{g: x_i(t_i + g) = R\}$ the time passed before his recovery. Note that, as infections can only occur during contacts, each t_i will be necessarily equal to the time of a contact. Let us define $\hat{r}_{ij}(t_i) = \min\{r: T_{ij}(r) > t_i\}$ the index of the first possible such contacts. Recovery of individuals will happen at a time $t_i + g_i$, where g_i follows a given recovery probability distribution with density $G(g_i)$. This parametrization is reversible, i.e. given t_i and g_i, it is easy to compute the state of $x_i(t)$ at any time t:

$$X(t_i, g_i, t) = \begin{cases} S & t_i < t \\ I & t_i \le t < t_i + g_i \\ R & t_i + g_i \le t \end{cases} \tag{1}$$

Since a node i has a finite probability to transmit the disease to a neighbor j in each of its contacts, one can compute the probability that the contagion will occur during the contact at time $T(\hat{r}_{ij}(t_i) + r_{ij}) \in (t_i, t_i + g_i)$, assuming that node j is still susceptible, simply as $R(r_{ij}) = \lambda(1 - \lambda)^{r_{ij}}$. Note that after time $t_i + g_i$, contagion will not take place because node i will be recovered. Given $\{g_k\}$ and $\{r_{ki}\}$, infection time t_i of a non-source node i must satisfy deterministically the condition $t_i = F_i(\{t_k\}, \{g_k\}, \{r_{ki}\})$ where

$$F_i(\{t_k\}, \{g_k\}, \{r_{ki}\}) = \min_{k \in \partial i: T_{ki}(\hat{r}_{ki}(t_k) + r_{ki}) < t_k + g_k} T_{ki}(\hat{r}_{ki}(t_k) + r_{ki}) \tag{2}$$

where ∂i denotes the set of neighbors of node i (i.e. nodes that share at least a contact with i) and it is conventionally assumed that the min is equal to $+\infty$ if the set is empty. Eq. (2) must hold because each neighbor k will be in the infected state in the time interval $[t_k, t_k + g_k)$, and will transmit at time $T_{ki}((\hat{r}_{ki}(t_k) + r_{ki}))$; the transmission that arrives first will be the one that succeeds. Suppose now an epidemic spreading is initiated by a spreader node i_0, which was infected at some time $-\varepsilon < 0$ before the first contact, which we conventionally fix at time $t = 0$. Our aim is to infer the initial spreader from just a single (possibly incomplete) observation of the state $\mathbf{x}(T)$ of the network at a later time T. The posterior distribution over the initial state of the networks can be easily written by means of Bayes theorem. In order to identify the initial spreader we should, in principle, maximize over the following posterior marginal probability:

$$i^* \in \arg\max_i \mathcal{P}(t_i = -\varepsilon | \mathbf{x}(T)) \tag{3}$$

we will give a very small prior probability γ to each initial seed, ensuring that configurations with more than one seed are overwhelmingly improbable (note that having at least one seed is necessary to explain any evidence of infection). The prior distribution $S(t_i)$ for the node i then reads:

$$S(t_i) = \gamma \delta[t_i; -\varepsilon] + (1 - \gamma)(1 - \delta[t_i; -\varepsilon]) \tag{4}$$

where $\delta[\cdot;\cdot]$ denotes the Kroenecker delta.

It is easy to see that (3) does not depend on ε. Indeed, for any $\varepsilon' > \varepsilon$, $\mathcal{P}(\mathbf{x}(T)|t_i = -\varepsilon') = \mathcal{P}(\mathbf{x}(T)|t_i = -\varepsilon|)(1 - \int_0^{\varepsilon' - \varepsilon} G(g)\, dg)$ implying that the two posteriors only differ by a constant factor, that has no relevance in (3).

The posterior distribution of infection times can now be written as $\mathcal{P}(\mathbf{t}|\mathbf{x}(T)) = \int_{\mathbb{R}_+^n} d\mathbf{g} \sum_{\mathbf{r}} \mathcal{P}(\mathbf{t}, \mathbf{g}, \mathbf{r}|\mathbf{x}(T))$ where

$$\begin{aligned} \mathcal{P}(\mathbf{t}, \mathbf{g}, \mathbf{r}|\mathbf{x}(T)) \;=\; & \frac{1}{Z} \prod_i \left(\delta[t_i; F_i(\{t_k\}, \{g_k\}, \{r_{ki}\})] + \delta[t_i; -\varepsilon]\right) \\ & \times \delta[X(t_i, g_i, T); x_i(T)]\, G_i(g_i) S(t_i) \prod_{ij} R(r_{ij}) \end{aligned} \tag{5}$$

The patient zero problem can be recast as the one of computing single site marginals $\mathcal{P}(x_i^0|\mathbf{x}^T)$ from the posterior distribution in equation (3). The problem of computing marginals over large dimensional probability distributions is in general intractable (NP-hard). In analogy with a previously introduced approximation method[6,7], we will tackle this problem by means of Belief Propagation, a method which is exact on acyclic graphs, and that was shown to perform very well on random and real contact networks in the discrete time scenario.

For each node in the network, the BP algorithm provides an estimate of the posterior probability that the node got infected at a certain time, and thus also the probability that the node was the origin of the epidemics.

Graphical model formulation. In order to apply BP, we will first formulate an alternative expression with no continuous variables, as the numerical representation of their distributions is problematic. Note that variables t_i are already discrete, as they live in the finite subset of the real line formed by all incoming contact times $H_i = \cup_{k\in\partial i, r=0,\dots n_{ki}} T_{ki}(r)$. Let us consider the ordered sequence of values $T_i(0) < \dots < T_i(n_i)$ in this subset, and define $T_i(n_i+1) = \infty$. To cope with continuous variables g_i, we will define from g_i a discrete variable $\tilde{g}_i \in \{0, \dots, n_i\}$, by exploiting the fact that $\sum_{\tilde{g}_i=0}^{n_i} \mathbb{I}[t_i + g_i \in [T_i(\tilde{g}_i),\ T_i(\tilde{g}_i+1))] = 1$ (here $\mathbb{I}$ is the indicator function of the condition in its argument), and that X and F are constant for g_i inside an interval $[T_i(\tilde{g}_i),\ T_i(\tilde{g}_i+1))$. We can recast (3) as

$$\begin{aligned}\mathcal{P}(\mathbf{t}|\mathbf{x}(T)) \;=\; & \sum_{\tilde{\mathbf{g}},\mathbf{r}} \frac{1}{Z} \prod_i \left(\delta[t_i;\, F_i(\{t_k\},\, \{T_k(\tilde{g}_k) - t_k\},\, \{r_{ki}\})] + \delta[t_i;\, -\varepsilon]\right) \\ & \times \delta[X(t_i,\, T_i(\tilde{g}_i) - t_i,\, T);\, x_i(T)]\, L_i(\tilde{g}_i,\, t_i) S(t_i) \prod_{ij} R(r_{ij})\end{aligned} \tag{6}$$

where $L_i(\tilde{g}_i,\, t_i) = \int_{T_i(\tilde{g}_i)-t_i}^{T_i(\tilde{g}_i+1)-t_i} G_i(g_i)\, dg_i$. In order to simplify the structure of the probability distribution factorization, we will introduce the variables $s_{ij} = S_{ij}(t_i,\, \tilde{g}_i,\, r_{ij})$ where

$$S_{ij}(t_i,\, \tilde{g}_i,\, r_{ij}) \stackrel{def}{=} \begin{cases} r_{ij} + \hat{r}_{ij}(t_i) & \text{if } \; T_{ij}(r_{ij} + \hat{r}_{ij}(t_i)) < T_i(\tilde{g}_i) \\ +\infty & \text{else} \end{cases} \tag{7}$$

The introduction of s_{ji} variables allows to simplify the structure of equation (6):

$$\mathcal{P}(\mathbf{t}|\mathbf{x}(T)) = \sum_{\tilde{\mathbf{g}}} \frac{1}{Z} \prod_i \psi_i(t_i,\, \tilde{g}_i,\, \{s_{ji}\},\, \{s_{ij}\}) \tag{8}$$

where

$$\begin{aligned}\psi_i(t_i,\, \tilde{g}_i,\, \{s_{ji}\},\, \{s_{ij}\}) \;=\; & (\delta[t_i;\, \min_{j\in\partial i} T_{ji}(s_{ji})] + \delta[t_i;\, -\varepsilon])\, \delta[X(t_i,\, T_i(\tilde{g}_i) - t_i,\, T);\, x_i(T)] \\ & \times L_i(\tilde{g}_i,\, t_i) S(t_i) \prod_{j\in\partial i} \sum_{r_{ij}} \delta[s_{ij};\, S_{ij}(t_i,\, \tilde{g}_i,\, r_{ij})]\, R(r_{ij})\end{aligned} \tag{9}$$

Belief Propagation equations. To briefly describe the essence of the BP method, let us consider a probability distribution over the variables $\underline{z} = \{z_i\}$ that has the following factorized form:

$$M(\underline{z}) = \frac{1}{Z} \prod_a F_a(\underline{z}_a) \tag{10}$$

where $\underline{z}_a$ is the subset of variables that F_a depends on. BP equations are a set of self-consistent relations linking the so-called *cavity messages* (or *beliefs*), a set of single-site probability distributions which are associated to each directed link in the graphical model defined by the joint distribution in equation (10). The general form of BP equations is the following:

$$p_{F_a\to i}(z_i) = \frac{1}{Z_{ai}} \sum_{\{z_j, j\in\partial a\setminus i\}} F_a(\{z_i\}_{i\in\partial a}) \prod_{j\in\partial a\setminus i} m_{j\to F_a}(z_j) \tag{11}$$

$$m_{i\to F_a}(z_i) = \frac{1}{Z_{ia}} \prod_{b\in\partial i\setminus a} p_{F_b\to i}(z_i) \tag{12}$$

$$m_i(z_i) = \frac{1}{Z_i} \prod_{b\in\partial i} p_{F_b\to i}(z_i) \tag{13}$$

where F_a is a *factor*, z_i is a variable, ∂a is the subset of indices of variables in factor F_a and ∂i is the subset of factors that depend on z_i. The terms Z_{ia}, Z_{ai} and Z_i are local partition function, serving as normalizations. To solve equations (11) and (12) an iterative procedure is typically used, where the cavity messages are initialized with homogeneous distributions and they are asynchronously updated until convergence to a fixed point[12,13]. While the computation of equation (12) is straightforward, the summation in (11) often involves a number of steps growing exponentially with the size of ∂a. In a number of interesting contexts, though, it is possible to devise efficient methods for computing this sum, and so reducing the computational complexity of the BP updates.

The inference problem of equation (8) is interpreted as a partial marginalization of a factorized distribution as in equation (10). In this settings, there are only two types of BP message, namely $m_{\psi_i\to(ij)}(s_{ij},\, s_{ji})$ and $m_{\psi_i\to i}(t_i)$,

and the corresponding updates are derived straightforwardly from equation (11). The node-to-factor BP messages, namely namely $m_{(ij)\to\psi_i}(s_{ij}, s_{ji})\, m_{i\to\psi_i}(t_i)$, can be computed very easily by virtue of equation (12).

Calling $M_i(\tilde{g}_i, t_i) = L_i(\tilde{g}_i, t_i) S(t_i)\, \delta[X(t_i, T_i(\tilde{g}_i) - t_i, T); x_i(T)]$, the BP equations for ψ_i are

$$\begin{aligned} p_{\psi_i\to(ij)}(s_{ij}, s_{ji}) \;\propto\; & \sum_{t_i}\sum_{\tilde{g}_i} M_i(\tilde{g}_i, t_i) \sum_{r_{ij}} \delta[s_{ij}; S_{ij}(t_i, \tilde{g}_i, r_{ij})] R(r_{ij}) \\ & \times \sum_{\{s_{ki}\}} \left(\delta[t_i; \min_{k\in\partial i} T_{ki}(s_{ki})] + \delta[t_i; -\varepsilon]\right) \prod_{k\in\partial i\setminus j} \\ & \times \sum_{r_{ik}} R_{ik}(r_{ik})\, m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \end{aligned} \tag{14}$$

$$\begin{aligned} = & \sum_{-\varepsilon<t_i<T_{ji}(s_{ji})} \sum_{\tilde{g}_i} M_i(\tilde{g}_i, t_i) \sum_{r_{ij}} \delta\big[s_{ij}; S_{ij}(t_i, \tilde{g}_i, r_{ij})\big] R(r_{ij}) \\ & \times \left\{ \prod_{k\in\partial i\setminus j} \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})\geq t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \right. \\ & \left. - \prod_{k\in\partial i\setminus j} \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})> t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \right\} \\ & + \sum_{t_i\in\{-\varepsilon, T_{ji}(s_{ji})\}} \sum_{\tilde{g}_i} M_i(\tilde{g}_i, t_i) \sum_{r_{ij}} \delta\big[s_{ij}; S_{ij}(t_i, \tilde{g}_i, r_{ij})\big] R(r_{ij}) \end{aligned} \tag{15}$$

$$\times \prod_{k\in\partial i\setminus j} \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})\geq T_{ji}(s_{ji})} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \tag{16}$$

where (15) follows because

$$\delta[t_i; \min_{k\in\partial i} T_{ki}(s_{ki})] = \prod_{k\in\partial i} \mathbb{I}[t_i \leq T_{ki}(s_{ki})] - \prod_{k\in\partial i} \mathbb{I}[t_i < T_{ki}(s_{ki})] \tag{17}$$

$$= \Big(\mathbb{I}\big[t_i < T_{ji}(s_{ji})\big] + \mathbb{I}\big[t_i = T_{ji}(s_{ji})\big]\Big) \left(\prod_{k\in\partial i} \mathbb{I}[t_i \leq T_{ki}(s_{ki})] - \prod_{k\in\partial i} \mathbb{I}[t_i < T_{ki}(s_{ki})] \right) \tag{18}$$

$$\begin{aligned} = & \mathbb{I}[t_i < T_{ji}(s_{ji})] \left(\prod_{k\in\partial i\setminus j} \mathbb{I}[t_i \leq T_{ki}(s_{ki})] - \prod_{k\in\partial i\setminus j} \mathbb{I}[t_i < T_{ki}(s_{ki})] \right) \\ & + \mathbb{I}[t_i = T_{ji}(s_{ji})] \prod_{k\in\partial\setminus i} \mathbb{I}[t_i \leq T_{ki}(s_{ki})] \end{aligned} \tag{19}$$

Similarly,

$$\begin{aligned} p_{\psi_i\to i}(t_i) \;\propto\; & \sum_{\tilde{g}_i} M_i(\tilde{g}_i, t_i) \times \left\{ \prod_{k\in\partial i} \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})\geq t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \right. \\ & \left. -(1 - \delta[t_i, -\varepsilon]) \prod_{k\in\partial i} \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})> t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \right\} \end{aligned} \tag{20}$$

Note that values $L_i(\tilde{g}_i, t_i) = \int_{T_i(\tilde{g}_i)-t_i}^{T_i(\tilde{g}_i+1)-t_i} G_i(g_i)\, dg_i$ can be pre-computed in a matrix $\mathcal{L}^i_{\tilde{g},\tilde{h}} \in \mathbb{R}^{(n_i+2)\times(n_i+2)}$ as each t_i must be equal $T_i(\tilde{h})$ for some $\tilde{h}$ in $\{0, \ldots, n_i+1\}$. The calculation of 16, 20 can be performed more efficiently by observing that all terms

$$Q^{\geq}_{ki}(t_i, \tilde{g}_i) = \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})\geq t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \tag{21}$$

$$Q^{>}_{ki}(t_i, \tilde{g}_i) = \sum_{r_{ik}} R_{ik}(r_{ik}) \sum_{s_{ki}: T_{ki}(s_{ki})> t_i} m_{(ki)\to\psi_i}(s_{ki}, S_{ik}(t_i, \tilde{g}_i, r_{ik})) \tag{22}$$

can be computed in time $O(n_i^2)$. Then equations (14)–(20) can be computed as

$$p_{\psi_i\to i}(t_i) = \frac{1}{z_i} \sum_{\tilde{g}_i} M_i(\tilde{g}_i, t_i) \left(\prod_{k\in\partial i} Q^{\geq}_{ki}(t_i, \tilde{g}_i) - \prod_{k\in\partial i} Q^{>}_{ki}(t_i, \tilde{g}_i)(1 - \delta[t_i, -\varepsilon]) \right) \tag{23}$$

$$p_{\psi_i\to(ij)}(s_{ij}, s_{ji}) = \sum_{-\varepsilon<t_i<T_{ji}(s_{ji})} m_{i\to\psi_i}(t_i)\sum_{\widetilde{g}_i} M_i(\widetilde{g}_i, t_i)\sum_{r_{ij}}\delta\Big[s_{ij}; S_{ij}(t_i, \widetilde{g}_i, r_{ij})\Big]R(r_{ij}) \times\left\{\prod_{k\in\partial i\setminus j} Q^{\geq}_{ki}(t_i, \widetilde{g}_i) - \prod_{k\in\partial i\setminus j} Q^{>}_{ki}(t_i, \widetilde{g}_i)\right\} + \sum_{t_i\in\{-\varepsilon, T_{ji}(s_{ji})\}} m_{i\to\psi_i}(t_i) \times\sum_{\widetilde{g}_i} M_i(\widetilde{g}_i, t_i)\sum_{r_{ij}}\delta\left[s_{ij}; S_{ij}(t_i, \widetilde{g}_i, r_{ij})\right]R(r_{ij})\prod_{k\in\partial i\setminus j} Q^{\geq}_{ki}(t_i, \widetilde{g}_i)\Bigg\} \tag{24}$$

on a loop over the possible values of t_i, $\widetilde{g}_i$.

The pseudocode in Algorithm 1 illustrates the implementation of our BP algorithm in detail. Please note that at each iteration we store factor-to-variable messages and marginals for each variable, the variable-to-factor messages being easily extracted at each iteration in view of the simple relation:

$$m_{i\to F_a} \propto \frac{m_i}{p_{F_a\to i}} \tag{25}$$

Algorithm 1 Belief Propagation based inference in continuous time SIR model.

input: A set of contacts over a graph G, a maximum tolerance $e_{\max}$ and a maximum number of iterations $\tau_{\max}$

output: A set of estimated marginals $\{m_i\}$

```
Initialize uniformly factor-to-variable messages {p}.
Initialize uniformly link marginals {m_(kl)} and node marginals {m_i}.
Initialize uniformly average marginals {m̄_i(t_i)}.
for τ = 1 to τ_max do
    Pick a random permutation π of nodes in G.
    e ← 0
    for j ∈ G do
        Pick node f = π(j).
        for k ∈ ∂f do
            Extract variable-to-factor messages m_(kf)→ψ_f ∝ m_(kf)/p_ψ_f→(kf).
        end for
        Extract variable-to-factor message m_f→ψ_f ∝ m_f/p_ψ_f→f.
        Update factor-to-variable message m_ψ_f→f = m_f via equation (23)
        Update average marginal m̄_j(t_j) using the recently computed m_f.
        for k ∈ ∂f do
            Compute outgoing factor-to-variable message m^new_ψ_f→(kf) via equations (24)
            e ← max{e, ‖m^new_ψ_f→(kf) − m_ψ_f→(kf)‖_∞}
            m_ψ_f→(kf) ← m^new_ψ_f→(kf)
            Update link marginal m_(kf) as m_(kf) ∝ m_(kf)→ψ_f p_ψ_f→(kf).
        end for
    end for
    if e < e_max then
        return marginals m_i = m_ψ_i→i.
    end if
end for
if τ = τ_max then
    return average marginals m̄_i(t_i).
end if
```

The time complexity of the computation of all messages $\{p_{\psi_i\to(ij)}\}_{j\in\partial i}$ exiting node i is $O(n_i^2\sum_{j\in\partial i} n_{ji}^2)$. Assuming a bounded number of contacts per individual n_i, this scaling results in an update with a number of operations per BP iteration which is linear in the number of edges (i.e. the number of pairs of individuals that are in contact) in the full system. On a BP fixed point, equation (13) is used to compute the marginal probability $m_i(t_i=-\varepsilon)$, which brings an estimation of the posterior probability $\mathcal{P}(t_i=-\varepsilon|\mathbf{x}(T))$ for the node to be the active before the first contact. Note that, in the present case, the marginal $m_i(t_i)$ is equal to the message $p_{\psi_i\to i}$. In the event that BP doesn't converge, reliable information can be extracted equivalently from the average value over iterations of the marginals $m_i(t_i)$, that we call $\overline{m}_i(t_i)$.

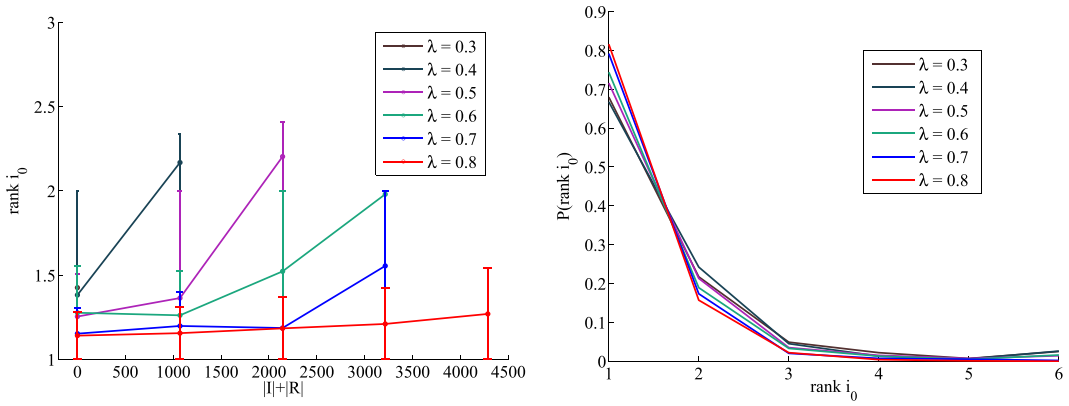


Figure 1. Left: Average absolute rank r_0 of the true patient zero as a function of the epidemic size $N_{IR} = |I| + |R|$ for $\mu = 0.5/year$ and increasing values of the infection probability λ in the network of sexual contacts. Each curve represents a sample of $M = 1000$ random instances. Lines are guide to the eye. Right: probability distribution of *rank* i_0 over the same epidemics for each value of λ.

Results

Patient zero detection in two large real contact time networks. We tested our methods on two large evolving networks: a database of time-stamped sexual interactions and a network of face-to-face contacts in a high school.

For each dataset, we simulated a large number of epidemic propagations, each one initiated from a unique random source (the patient zero, or seed). The nodes in the network are then ranked in decreasing order of their estimated posterior probability of being the origin of the observed epidemics: the position of the true origin in the ranking provided by the algorithm is a good measure of the efficacy of the method. In what follows, i_0 stands for the index of the true origin of the epidemics, its rank being indicated by $r_0 = rank(i_0)$. For simplicity, we will always consider homogeneous epidemic parameters $\lambda_{ij} = \lambda$ and $G(g_i)$ an exponential distribution with rate constant μ.

The first dataset comes from a database of sexual encounters between clients and escorts on a Brazilian website, covering the beginning of the community, September 2002 through October 2008, and composed of a total of $E = 50185$ contacts between between 6642 escorts and 10106 sex-buyers. This kind of data are particularly relevant in the study of spreading of Sexually Transmitted Infections (STI), and have been previously used to model the diffusion of HIV by means of simple SI-SIR compartmental models[11]. We build a bipartite evolving network, focusing on the last two available years of operation of the website ($E = 29628$ contacts, slightly over half of the dataset) in order to skip the initial period where reporting of encounters are very sparse and incomplete. For each value of λ, we simulate $M = 1000$ single source epidemic propagations, with a recovery rate equal to $\mu = 0.5/year$. Concerning algorithmic efficiency, the efficient implementation of the BP updates presented in the previous section allows us to perform inference in large-scale contact network with a remarkably small computational cost. As an example, a parallel C++ implementation with 5 concurrent processes took roughly 11 minutes to converge to a solution with $rank(i_0) = 1$ (perfect identification of the source) on a single instance with an epidemic size of $N_{IR} = 1810$ individuals, obtained with $\lambda = 0.4$ and $\mu = 0.5/year$.

In Fig. 1 we show the average absolute rank r_0 of the true first infected individual i_0 as a function of the epidemic size $N_{IR} = |I| + |R|$ (i.e. the number of infected and recovered sites), whose values are discretized in intervals of width equal to 0.1. The low values of the rank show the effectiveness of the method.

The second dataset consists of a collection of Close Proximity Interactions (CPIs) obtained by means of wireless sensor network technology (TelosB motes)[8]. Data were collected in a US high school and provide an almost complete account of face-to-face interactions during a whole day at school. All in all 798 individuals were monitored, corresponding to the 94% of the total school community, and 2148991 unique Close Proximity Records (CPR) were acquired. A single CPR corresponds to a close proximity detection event between two motes (max. 3 meters). The authors of the study perform an aggregation of the raw data in *interactions*, defined as continuous sequences of CPRs between the same two nodes. Our choice was to go back to the raw data and investigate the spreading process at the level of single CPRs, using the intensity signal as a proxy for the closeness of a face-to-face contact (a detailed account is present in Salathe *et al.* - Supplementary Information[8]). We constructed a set of evolving networks by setting a threshold θ_{int} for the signal intensity of the motes, thus resulting in denser networks for smaller θ_{int}, where more weaker (and distant) contacts are taken into account. A second interesting possibility could be to allow for a probability of contagion that depends on the proximity (i.e. the strength of the signal). We did not persue this route as the dependence of probability cf contagion on the distance is hard to determine (to our knowledge, no study provides this information for known diseases), and moreover we don't have information on the correspondence between signal strength and distance. Besides, a threshold-like dependence on the distance could be adequate for contagion of many infectious diseases, such as non-airborne ones. In any case, the analysis technique presented here can be adapted directly to the case of contact-dependent probability would the needed modelling information mentioned above be available in the future.

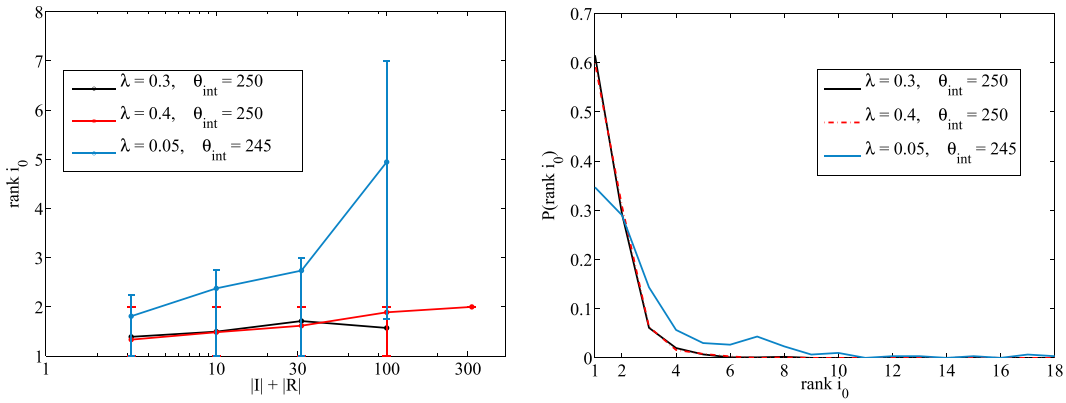


Figure 2. Left: Average abolute $r_0 = rank(i_0)$ as a function of the total epidemic size $N_{IR} = |I| + |R|$ in the network of face-to-face contacts in a high school for different values of the threshold θ_{int}. Each curve is an average over $M = 1000$ ($\theta_{int} = 250$) or $M = 300$ ($\theta_{int} = 245$) random instances. Lines are guide to the eye. Right: probability distribution of $rank i_0$ over same epidemics for the three set of parameters.

Three representative examples are show in Fig. 2, which displays the average rank of the true first infected individual i_0 for different values of λ and threshold θ_{int}. Two values of $\lambda = 0.3, 0.4$ are explored for $\theta_{int} = 250$. Then we attempted with the much denser graph resulting of considering $\theta_{int} = 245$. Here the infection probability λ has been chosen to maintain the same average number of infections as the case $\theta_{int} = 250$, $\lambda = 0.3$.

Reconstruction of causality. Consider the problem of inferring for each non-susceptible individual i at time T, the individual k from which he contracted the infection. The probability p_{ktoi} of such transmission event corresponds to

$$p_{k\to i} = \sum_{\tilde{r}_{ki}} \mathcal{P}(t_i = T(\hat{r}_{ki}(t_k) + \tilde{r}_{ki}),\ r_{ki} = \tilde{r}_{ki} | \mathbf{x}(T)) \tag{26}$$

Once $p_{k\to i}$ has been computed for every (ordered) pair (ki), a prediction will be formed by the subset of pairs with probability larger than a given threshold. A *receiver operating charasteristic* (ROC) curve can be computed by considering the performance for all possible thresholds. The ROC curve for an instance of an outbreak on the sexual contacts dataset is shown in Fig. 3, along with the inferred pairs in one single point of the curve. It is evident that a large fraction of the entire history of the propagation can be reconstructed with a high degree of reliability, despite the apparently limited amount of information available in a single observation of the nodes' states.

For a comparison between the true infection times t_i^{true} of each node and the ones that can be inferred by our method, we show in the left panel of Fig. 4 a scatter plot of t_i^{true} versus $\bar{t}_i$ for a single epidemic cascade in the network of sexual contacts, with epidemic parameters $\lambda = 0.4$ and $\mu = 0.5/year$ (in the resulting epidemics the number of infected individuals is $|I| = 991$, total epidemic size being $N_{IR} = 1070$). The infection times $\bar{t}_i$ have been simply obtained by averaging over the marginal posterior distribution $\mathcal{P}(t_i|\mathbf{x}(T))$ of single-node infection times. In addition, we show in the right panel of Fig. 4 the Average Time Error (*ATE*) between $\mathbf{t}^{true}$ and $\bar{\mathbf{t}}$, which we define as $ATE = \sum_{i\in G} |\bar{t}_t - t_i^{true}|/(|I| + |R|)$, for a set of 200 different samples of simulated epidemic outbreaks in the same network and with the same epidemic parameters as in the previous example. The reconstruction of the dynamical history is remarkably good over a wide range of epidemic sizes.

Partial and noisy observations. It is not difficult to extend the present model to account for observations affected by some kind of uncertainty. One simply introduces the Observational Transition Matrix (OTM) $o_i(y_i|x_i)$, containing the transition probabilities from the true state x_i to the observed state: in order to perform inference, one has to some over all the possible true unobserved states with a weight given by the corresponding entry in the OTM, a task which is easily accomplished in the BP inference scheme. The identity matrix $o_i(y_i|x_i) = \delta[y_i; x_i]$ corresponds to the a case in which no noise enters the observations. Please note that, in this generalized scheme, we can simply take into account partial observability with a totally flat OTM $o_i(y_i|x_i) = \frac{1}{3}$ for unobserved nodes.

Firstly, we consider the case of partial observations, i.e. the case in which only a subset of nodes are accessible for observation at time T: this is the standard realistic scenario in practical applications, when a complete monitoring of a full network is infeasible in the general case. We simulate a number of epidemic spreading in the contact network and model the partial observability by a fixed probability p_{ob} of observing a node at time T. In what follows, we will use the sexual encounters dataset, mostly because of its epidemiological relevance.

Results for decreasing values of p_{ob} in the network of sexual contacts are shown in the left panel of Fig. 5 (the complete observation case $p_{ob} = 1$ is shown in the dashed line for reference). The BP method happens to be highly resilient even with a high amount of hidden information at time of observation T.

Turning to the problem of uncertain observations, we use a very simple symmetric model for observational noise. Let us consider the following OTM:

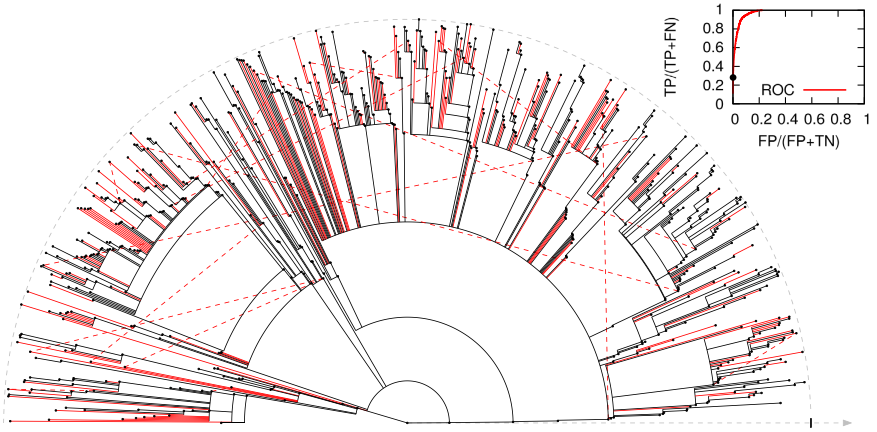


Figure 3. Reconstruction of the causal history of transmissions on a simulated epidemic outbreak on a sexual contact network infecting 719 individuals (black dots) from a snapshot of the infection state at time 2y (corresponding to the outer dashed semicircle). The figure represents the tree of transmissions in the outbreak originated in the central node (time flows in the outwards radial direction). Radial segments correspond to 718 true transmission events. The 4130 oriented pairs of infected individuals that were in contact were ordered by their decreasing estimated posterior probability of transmission. A ROC curve (area equal to 0.898) of the ordered list is shown in the top right panel, with the black point corresponding to the main figure. In the main figure, red full radial segments correspond to 202 correctly inferred transmissions (true positives, TP), black ones to non-inferred transmissions (false negatives, FN) and red dashed lines to the 16 wrongly predicted transmissions (false positives, FP).

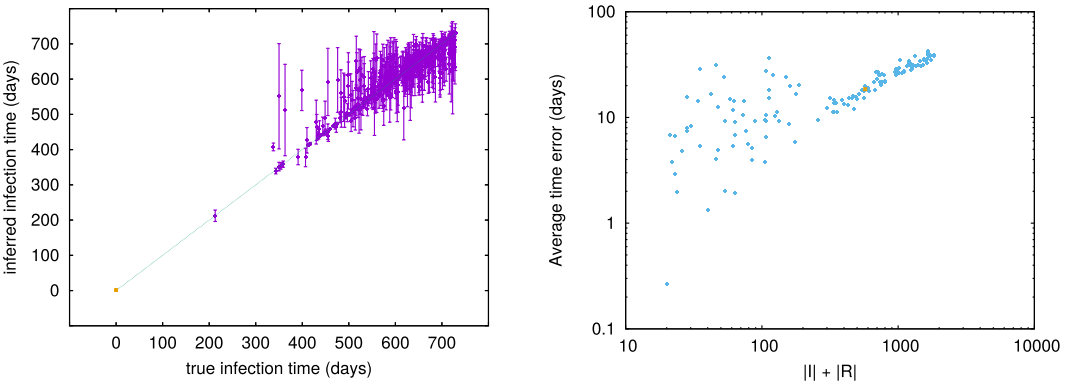


Figure 4. Left panel: inferred vs true infection time for a single epidemic cascade in the network of sexual contacts. Epidemic parameters are $\lambda = 0.4$ and $\mu = 0.5/year$. The vertical bars represent the standard deviation of the inferred infection time computed from the BP marginals. The zero patient, highlighted in orange, is correctly identified as the first infected individual. Right panel: Average Time Error (*ATE*) sorted as a function of the epidemic size N_{IR} in 140 samples (we simulated 200 epidemic cascades and then focused on samples with $N_{IR} > 20$, where enough information is present at the time of observation). Epidemic parameters are the same as in the left plot, each point represents the *ATE* value for a single epidemic cascade. The orange dot corresponds to the cascade on the left panel.

$$o_i(y_i|x_i) = (1 - \nu)\delta[y_i; x_i] + \frac{\nu}{2}(1 - \delta[y_i; x_i]) \tag{27}$$

This matrix describes a kind of symmetric noise, where a node that is in the state x has a probability $1 - \nu$ of being correctly observed in its state, and a probability ν of being observed incorrectly in one of the other two states. Suppose, for example, node i is S (susceptible) at observation time T: for a given noise probability ν, there is an equal probability $\frac{\nu}{2}$ for the node i to be observed in the R (recovered) or I (infected) state - the same holds for the others two cases, equivalently. In Fig. 5, right panel, we show that our BP algorithm is highly robust even to a significant amount of noise, up to 30%.

Discretization and binning. In order to ascertain the eventual loss of inference precision due to time-discretization, we performed the following experiment on the sexual intercourse dataset[10]. We generated $M = 100$ random epidemics with the semi-continuous time model with instantaneous probability of transmission λ.

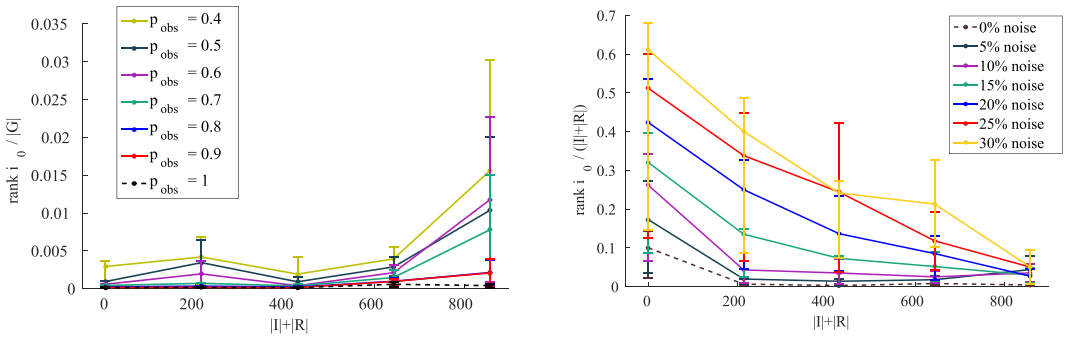


Figure 5. Inference performance with partial or noisy observations in the network of sexual contacts. In each curve we show the average normalized rank over $M = 100$ random instances of the true patient zero i_0 as a function of the total epidemic size $N_{IR} = |I| + |R|$ for $\lambda = 0.2$, $\mu = 0.5/year$. Left panel: $r_0 = rank(i_0)$ normalized over $|G|$ vs N_{IR} for different values of probability of observation p_{obs}; dashed curve is the case with full observations. Right panel: normalized $r_0 = rank(i_0)$ vs N_{IR} for different noise intensity ν; dashed curve is the case with no noise. Lines are guide to the eye.

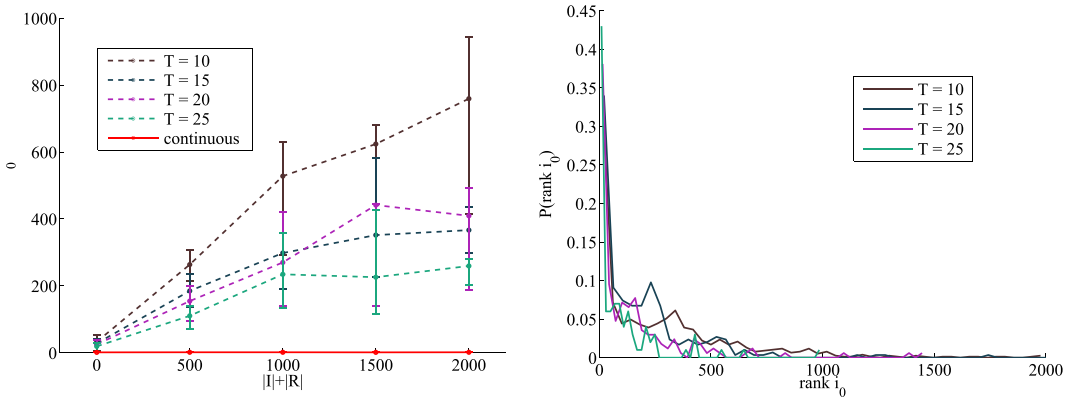


Figure 6. Left: Comparison between the continuous time method and the discretized version in the network of sexual contacts for infection probability $\lambda = 0.4$ and $\mu = 0.5/year$. Each curve is the average absolute rank r_0 over $M = 100$ samples as a function of the epidemic size N_{IR}. Full curve: continuous time version. Dashed curves: contacts have been aggregated in effective contacts so that final time of observation T is systematically increased from $T = 10$ to $T = 25$. Lines are guide to the eye. Right: probability distribution of *rank* i_0 over the same epidemics for the four values of T. The distribution for the continuous case has been omitted since probability is extremely concentrated around $r_0 = 1$ and appears off-scale.

Separately, for each value of $T = 10, 15, 20, 25$ we produced a discrete time temporal network from by sub-dividing the time interval $[t_0, t_1]$ into T equal subintervals (time-steps) of length $\Delta = (t_1 - t_0)/T$, aggregating all contacts falling in a given interval $[t\Delta; (t+1)\Delta]$ into a time-step dependent weight $\lambda_{ij}^t = 1 - (1 - \lambda)^{k_t}$, where k_t is the number of contacts in the interval[8,6].

We then performed the inference analysis on the M previously generated epidemics both in the discretized network with the the discrete-time BP algorithm[6] and using the semi-continuous time inference BP algorithm presented here. We investigated the difference in performance for different combinations of the epidemic parameters, noting that the semi-continuous time method strikingly outperforms the discretization procedure in all cases. As an example, we show in Fig. 6 a comparison of the continuous time method versus the discretized version for $\lambda = 0.4$ and $\mu = 0.5/year$ with an increasing number of time bins.

Discussion

In this work, we developed an inference framework to analyze the dynamics of infection in temporal contact networks with continuous or very fine-grained time resolution. We showed by means of simulations on real contact networks how the approach is able to reconstruct with great degree of accuracy both the source of the epidemics, the infection times and the underlying epidemic causal history from the mere observation of the state of the system (noisy or incomplete) at a single instant in a wide range of parameters.

Moreover, we were able to quantify the loss of information due to time-discretization, demonstrating a remarkable improvement with continuous time inference when compared with time discretized data even for a relatively large number of time sub-intervals.

It would be interesting to apply this technique on other relatively closed communities where the interactions can be monitored but the infections themselves are hidden, such as in hospital wards[14] and for applications to computer virus forensics[4]. The ability to reconstruct the epidemic history and causality of transmissions could prove to be helpful to devise better containment strategies. In those cases, generalizations of the method to epidemic models related to SIR such as SEIR and other distributions of recovery time different from exponential could be necessary.

References

1. Antulov-Fantulin, N., Lancic, A., Stefancic, H., Sikic, M. & Smuc, T. Statistical Inference Framework for Source Detection of Contagion Processes on Arbitrary Network Structures. In *2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems Workshops (SASOW),* 78–83 (2014).
2. Lokhov, A. Y., Mézard, M., Ohta, H. & Zdeborová, L. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* **90,** 012801 (2014).
3. Pinto, P. C., Thiran, P. & Vetterli, M. Locating the source of diffusion in large-scale networks. *Physical Review Letters* **109,** 068702 (2012).
4. Shah, D. & Zaman, T. Detecting sources of computer viruses in networks: theory and experiment. *ACM SIGMETRICS Performance Evaluation Review* **38,** 203–214 (2010).
5. Shah, D. & Zaman, T. Rumors in a network: Who's the culprit? *Information Theory, IEEE Transactions on* **57,** 5163–5181 (2011).
6. Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A. & Zecchina, R. Bayesian inference of epidemics on networks via belief propagation. *Physical Review Letters* **112,** 118701 (2014).
7. Altarelli, F., Braunstein, A., Dall'Asta, L., Ingrosso, A. & Zecchina, R. The patient-zero problem with noisy observations. *Journal of Statistical Mechanics: Theory and Experiment* **2014,** P10016 (2014).
8. Salathé, M. *et al.* A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* **107,** 22020–22025 (2010).
9. Isella, L. *et al.* What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* **271,** 166–180 (2011).
10. Rocha, L. E. C., Liljeros, F. & Holme, P. Information dynamics shape the sexual networks of internet-mediated prostitution. *Proceedings of the National Academy of Sciences* **107,** 5706–5711 (2010).
11. Rocha, L. E. C., Liljeros, F. & Holme, P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol* **7,** e1001109 (2011).
12. Yedidia, J. S., Freeman, W. T. & Weiss, Y. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems* **13** (2001).
13. Mézard, M. & Montanari, A. *Information, Physics, and Computation* (Oxford University Press, 2009).
14. Vanhems, P. *et al.* Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE* **8,** e73970 (2013).

Acknowledgements

AB acknowledges support by Fondazione CRT under the initiative "La Ricerca dei Talenti".

Author Contributions

A.B. and A.I. conceived the experiments, conducted the experiments and analyzed the results. All authors reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Braunstein, A. and Ingrosso, A. Inference of causality in epidemics on temporal contact networks. *Sci. Rep.* **6**, 27538; doi: 10.1038/srep27538 (2016).